Administrative Data Initiatives at Statistics Canada

Julie Trépanier, Jean Pignal and Don Royce

Statistics Canada
170 Tunney's Pasture Driveway, Ottawa (Ontario), K1A 0T6, Canada

Introduction

Statistics Canada has long been using administrative data (i.e., information that is collected by organizations and departments for their own purposes) in its statistical programs and is determined to increase the use of such data when it leads to a better outcome - that is a better balance between relevance, quality, costs and respondent burden. To achieve this objective, the organization created an Administrative Data Secretariat (ADS) in the fall of 2012, with the mandate to develop and implement a corporate approach to increasing the use of administrative data.

The Secretariat has undertaken several initiatives as part of this mandate. They include: a review of the legal, policy and organizational frameworks for the statistical use of administrative data that exist in Canada and elsewhere, to identify approaches that might be adopted at Statistics Canada; the construction of a central inventory of administrative data sources currently received by Statistics Canada, to understand better what data the organization uses and how it could use it better; and the launch of a Census Program research project, to study the feasibility of building a statistical database of all Canadians and their basic demographic information by using multiple administrative data sources.

This paper will first provide an overview of how the use of administrative data at Statistics Canada has evolved, leading to the creation of the ADS. This will be followed by a description of the three initiatives mentioned above.

Use of administrative data over time

Statistics Canada has been using administrative data for nearly a century. In 1918, the *Statistics Act* created the Dominion Bureau of Statistics, now known as Statistics Canada, a national statistics institute (NSI) with broad powers to collect administrative and survey data for statistical purposes. One outcome of this new legislation was the transmission by 1921 of vital statistics records from all provinces that were part of Canada at the time. Since those early days, the use of administrative data at Statistics Canada has continued to expand. In addition to vital statistics, administrative data such as those on international trade, health, justice and education are used directly in a variety of statistical programs. In fact Section 3(d) of the current *Statistics Act* mandates Statistics Canada to promote the avoidance of duplication in the information collected by departments of government.

As more and more sample surveys were launched in the second half of the 20th century, the need for survey frames that allow the design of more efficient surveys increased. As a result, the Business Register was developed in the 1980s based on data from the Canada Revenue Agency (CRA). In the 1990s, the first Address Register, which relied on the T1 Personal Income Tax File from the CRA, municipal assessment rolls, telephone and electricity billing files, was put in place to serve as a coverage improvement tool for the 1991, 1996 and 2001 Censuses of Population. As more and more administrative data sources were used to maintain the Address Register, its quality improved considerably. As a result, the Address Register became the dwelling frame for a large portion of the 2006 and 2011 Censuses of Population, allowing the mail-out of letters or questionnaires to 80% of the private dwellings in Canada in 2011.

In the last two decades, reducing response burden by using administrative data to partially replace survey data has become increasingly important. On the business statistics side, the monthly Survey of Employment, Payrolls and Hours has based its estimates of total number of employees and gross monthly payroll on the same two variables collected by CRA on the payroll deduction accounts forms since 1994. It has also used the administrative variables total number of employees and gross monthly payroll to strengthen the production of other survey estimates via calibration-type estimation methods. The annual business surveys program started using income tax data from CRA to estimate for the very small businesses in 1997 and to later reduce the size of the sample of simple businesses that need to be sent to collection. In a similar way, key monthly business surveys (manufacturing, retail and wholesale, food services) started using Good and Services Tax sales collected by CRA in 2004-2005 to reduce the number of units sent to collection. On the social statistics side, personal income tax data started being used to replace revenue questions collected on the Survey of Labour and Income Dynamics in 1995. This

approach was extended to the Census of Population Program in 2006, the Survey of Household Spending in 2010 and the Longitudinal and International Survey of Adults in 2011.

Finally, as more and more administrative data are acquired, new analytical opportunities arise. A notable example is the linkage of data on permanent immigrants, obtained since 1980 from Citizenship and Immigration Canada, with taxation data, obtained from CRA since 1982. The result is the Longitudinal Immigration Database (IMDB), a comprehensive source of data on the economic behaviour of the immigrant taxfiler population in Canada. It constitutes the only source of data that provides a direct link between immigration policy levers and the economic performance of immigrants.

Creation of the Administrative Data Secretariat

Despite this long history of using administrative data in various ways, there are many reasons for Statistics Canada to continue to look for other opportunities. The Government of Canada has made commitments to reduce the reporting burden on Canadian businesses and as such, Statistics Canada has made its own commitments to reduce the time spent by businesses to complete its surveys by using existing information whenever possible. Statistics Canada also seeks opportunities to use existing sources of information as a more cost-effective alternative to developing new surveys with their attendant collections costs. The increasing difficulties in reaching household survey respondents has also been identified as a corporate risk and using administrative data is one of the mitigation strategies. These pressures, coupled with the opportunities created by new administrative data sources and the increased ability to process and use administrative data, have pushed Statistics Canada to take a closer look at how it obtains and uses administrative data.

Several different committees at Statistics Canada concluded that better coordination was needed to increase the use of administrative data, and that a designated group should be created to further this goal. With the exception of the Tax Data Division, which coordinates the acquisition and processing of the data coming from CRA, administrative data activities are largely decentralized at Statistics Canada and are normally managed by the most relevant subject matter division (e.g., vital statistics under the Health Statistics Division).

The considerations above led to the creation of the Administrative Data Secretariat in September 2012. The mandate of the ADS was not to take over all activities surrounding the acquisition, use and management of administrative data at Statistics Canada but rather to develop and implement a corporate approach to increase the use of administrative data over a two year period (April 1, 2013 to March 31, 2015). Three full-time equivalent employees were allocated to the ADS for each of the two years. This mandate itself consisted of three main objectives:

1. Put in place a governance structure, i.e., policies, directives, guidelines, practices and tools that will support statistical programs at Statistics Canada in the acquisition, management and efficient use of administrative data.

2. Launch initiatives that seek to optimize the methods and processes surrounding administrative data.

3. Provide support to statistical programs in their research of new sources of administrative data and related methods.

To pursue these objectives, several different activities were launched. A description of a subset of these follows.

International review of legal, policy and organizational frameworks for the statistical use of administrative data

Under the first objective presented above, the ADS conducted a review of international frameworks for the statistical use of administrative data (Royce 2013). The study consisted of a comparison of the legal, policy and organizational frameworks in five countries that have a similar statistical environment to Canada: Ireland, the United Kingdom, New Zealand, Australia and the United States. It also looked at practices and approaches disseminated by several international statistical organizations, namely the United Nations, the United Nations Economic Commission for Europe (of which Canada is a member) and its Conference of European Statisticians, Eurostat and the European Statistical System.

Legal frameworks

The review underlined the fact that a country's statistical use of administrative data is fundamentally influenced by the legislative environment within which the NSI functions. While the legal basis for protection of confidentiality and privacy was quite consistent across countries, considerable variation was found across NSIs in their legislative authority to influence, access and use administrative data for statistical purposes, with Canada having neither the strongest nor the weakest legislation. For example, Section 13 of the *Statistics Act* gives Statistics Canada the authority to access virtually any administrative record from

any government department, municipal office, corporation, business or organization to fulfill the purposes of the Act; by comparison, the United Kingdom must obtain Parliament's authority to access any new administrative data source, and this access is subject to the agreement of the supplying department.

Of the countries reviewed, the Central Statistics Office (CSO) of Ireland has the most comprehensive legislative authority for the statistical use of administrative data, particularly for information held by public authorities. Of interest are the following measures in Sections 30 and 31 of the Irish *Statistics Act 1993*:

- An explicit statement to the effect that the statistical legislation overrides other enactments (with a few natural exceptions, e.g., national security);
- A legal mechanism to arbitrate between the CSO and the custodial organization concerning access to data;
- An explicit statement that data held by other government departments are to be provided to the CSO free of charge;
- A requirement that other departments of government cooperate with the CSO in examining the statistical potential of the administrative records held by those organizations;
- A requirement that other departments of government consult with the CSO when creating or redeveloping their administrative records system, and that they accept reasonable recommendations that would improve the usability of these records for statistical purposes.

Policy frameworks

The review of policy frameworks looked at the overall government environment that promotes or constrains the statistical use of administrative data. Such an environment can take the form of a statistical code of practice, of which the statistical use of administrative data is one part, as well as government-wide policies and initiatives on privacy, information management and sharing of information among government departments.

The United Kingdom, Australia, New Zealand and the United States all have some version of a national code of statistical practice, and Ireland is developing such a code for 2014. Eurostat also has a code of statistical practice that applies to European Statistical System members. These codes of practice generally include principles and protocols that explicitly address the statistical use of administrative data, for example:

- Administrative data are viewed as a strategic asset for research and statistical purposes as well as for administrative purposes;
- Direct data collection should only be carried out when the information requirements cannot be met from existing data;
- Statistical authorities should be involved in the design of administrative record systems;
- Statistical authorities should be involved in assuring data quality of the administrative sources;
- Administrative authorities need to consult with statistical authorities before making changes that could affect the statistical use of the data;
- The administrative and statistical uses of data should be kept functionally separate by legal, policy and organizational safeguards; and
- The statistical use of administrative data should be transparent.

The review also found that, compared to Canada, the other countries examined appear to have taken a more government-wide approach to statistical policy development and related initiatives where the statistical use of administrative data is concerned. While this might be expected in decentralized statistical systems such as those in the United Kingdom and the United States, it also appears to be the case in the more centralized statistical systems in Australia, New Zealand and Ireland.

For example, some countries have established one or more cross-government statistical liaison groups as a forum for statistical issues of common interest, including the statistical use of administrative data. The activities of such groups can include the development of codes of practice such as those described above; the sponsoring and overseeing of studies that identify administrative data with statistical potential and/or changes to administrative systems that would make them more useful for statistical purposes; developing government-wide approaches to statistical data integration; and promoting common data quality frameworks and tools.

Several NSIs have recognized the potential of linking administrative data to survey data or other administrative data, or linking administrative data longitudinally and have promoted the concept of "statistical data integration". For example, Australia has established government-wide principles and approaches for statistical data integration, including a formal accreditation process for organizations undertaking so-called "high risk" data integration activity. Statistics New Zealand has been designated as the government custodian for statistical integrated datasets and is developing an Integrated Data Infrastructure to bring together

linked data on individuals and businesses. The CSO of Ireland is developing a Person Activity Register with similar goals and approaches.

Finally, several countries have advisory or coordinating bodies that not only provide advice to the NSI, but to the government itself. This is the case for: the Irish National Statistics Board; the UK Statistics Authority; the Australian Statistics Advisory Council; and the Office of Management and Budget, the Government Accountability Office, and the Committee on National Statistics of the National Research Council of the National Academies in the United States.

Organizational frameworks

Organizational frameworks apply at the level of the NSI, and typically include: the organizational arrangements between the NSI and the organizations supplying the data; the organizational structure within the NSI for receiving, using and managing the administrative data; and the NSI's written policies, directives, protocols, standards, guidelines, etc., that govern the statistical use of administrative data. Although the specifics of organizational frameworks differ across NSIs, three common themes emerged from the review.

1. NSIs need to establish effective processes and structures for identifying, influencing and accessing administrative data, and for ensuring that data are of sufficient quality for statistical purposes.
2. NSIs need to create the proper data stewardship environment to ensure that administrative data are treated with proper care and in accordance with legal requirements.
3. NSIs must ensure long-term stakeholder support by being transparent about what administrative data are used and how they are used.

In regards to effective processes, all NSIs reviewed recognize the need to cultivate working relations with the administrative data suppliers, and use mechanisms such as Memoranda of Understanding, Service Level Agreements, bilateral liaison committees, working groups, *quid pro quo* services, and so on. Concerning the organization of administrative data activities within NSIs, the degree of centralization for receiving, processing and using administrative data varies, but there does appear to be some trend towards centralizing at least the reception and management of administrative data. Concerning data quality, all of the countries examined recognize the special challenges of using administrative data for statistical purposes, and many NSIs and international organizations have developed or are developing data quality frameworks, guidelines, processes, and other tools to help assess the quality of administrative data for statistical purposes.

The U. S. Census Bureau has developed an "Administrative Records Handbook", a good example of data stewardship that serves as a guide for its managers to the policies, procedures and practices relevant to using administrative data in statistical programs. The Census Bureau has also developed an Administrative Records Tracking System (ARTS) to document and control all uses and users of administrative data. The ARTS serves as a central repository for metadata on the administrative datasets, for agreements with data providers, and for documenting the project review and approval process. As noted earlier, several countries have embraced the concept of statistical data integration, and all countries examined have some form of internal policy on record linkage or data integration. Statistics New Zealand appears to have the most comprehensive Data Integration Policy, supported by a Data Integration Manual that goes into methodological and operational details.

Concerning transparency with stakeholders, several of the countries examined publicly identify the administrative sources they use to produce statistics, and/or the arrangements with major data suppliers. For example, the Office for National Statistics, like other statistical organizations in the United Kingdom, publishes a Statement of Administrative Sources. Australia identifies important administrative sources as part of its list of Essential Statistical Assets, and the Central Statistics Office of Ireland posts its Memoranda of Understanding with the Revenue Commissioners on its website. Some NSIs also have mechanisms for important stakeholders to be consulted when data integration projects are being considered, and most countries post approved data integration projects on their websites.

Administrative Data Inventory

Even before the ADS was created, Statistics Canada noted it did not have in place a central repository of information about administrative datasets coming to Statistics Canada from other organizations. In the summer and fall of 2012, over 40 statistical programs within Statistics Canada were asked to provide metadata about the datasets they receive. The information requested was about the supplying organization, the nature of the file, the nature of the underlying agreement with the provider, the statistical program using it and how the data are used (e.g., for survey frame, edit and imputation, direct tabulation). A dataset may appear more than once in the inventory if it is used by more than one statistical program.

This initial effort resulted in a first inventory of administrative datasets received during fiscal year 2012-2013. Highlights of this first version are presented below in Table 1.

Table 1: Administrative Datasets Collected by Statistics Canada in Fiscal Year 2012-2013, by Source and Usage[1]

Source of Dataset	Number of Datasets	Usage by Statistical Program at Statistics Canada[2]				
		Economic and Environmental	Socio-economic	Censuses	Statistical Infrastructure	Cost-recovery
Federal	187	148	51	3	37	2
Provincial/Territorial /Municipal	135	97	14	2	22	0
Private	166	140	1	2	18	5
Foreign	11	11	0	0	0	0
Others	13	2	6	2	3	0
Total	**512**	**398**	**72**	**9**	**80**	**7**

[1] These results are preliminary. Statistics Canada is currently reviewing the way the information on administrative datasets is consolidated. Some duplication in the inventory is suspected due to the lack of a common naming convention.
[2] There can be multiple users of one dataset.

Table 1 shows that most datasets come from the federal government. This was expected since it remains easier to negotiate an agreement to acquire and use a dataset from a single federal department than to negotiate agreements with 13 provincial and territorial jurisdictions or from a large number of municipalities. This translates into greater uses of administrative data in Statistics Canada's Economic and Environmental Statistics Program, because many of the administrative programs that could be useful to the Socio-Economic and Censuses Statistics Programs are under provincial jurisdiction (e.g., health, education, justice).

Although very rudimentary, this first inventory has led to a better understanding of Statistics Canada's administrative data holdings. The inventory has been made available internally to all employees to encourage statistical programs to improve the quality of the information for the next iteration (fiscal year 2013-2014), to increase the use of sources that Statistics Canada already has and to identify opportunities for optimization. For the latter, the Administrative Data Inventory allows the identification of files that are used by more than one statistical program and describes how the files are used. The ADS has started investigating whether there would be room for optimization in methods and processes for frequently used files.

Research into a Canadian Statistical Demographic Database

The Canadian Statistical Demographic Database (CSDD) is a research project that is examining the extent to which administrative data from various sources could be used to create an up-to-date database of the resident population of Canada and their usual place of residence, along with basic demographic information such as sex and date of birth. While no other country has successfully implemented an administrative Census of this kind, the United Kingdom and New Zealand are both investigating a similar paradigm. The UK is studying such options for its 2021 Census, while New Zealand plans to complete the evaluation of such options and the implementation of the option deemed feasible in the 2026-2031 timeframe. The CSDD is managed by the ADS on behalf of the 2016 Census Program. This work receives separate funding but falls under the third objective of the ADS presented earlier.

The CSDD will be initially constructed by updating previous Census information with a variety of administrative files, such as records of births and deaths from the 13 provinces and territories of Canada, T1 income tax filer information from the Canada Revenue Agency, Citizenship and Immigration files on permanent and temporary residents and the Indian Register, which enumerates Registered Indians in Canada and is maintained by the Aboriginal Affairs and Northern Development Canada.

The initial database will be constructed with a reference date of May 2011, by starting with the 2006 Census of Population and chronologically updating the database through matches with appropriate administrative sources, until all population movements (i.e., natural growth, international migration and migration within Canada) up to May 2011 are accounted for. This file will then be compared to the 2011 Census of Population in order to assess the coverage and quality of the underlying data. At this point, gaps and weaknesses (e.g., individuals not covered by administrative sources) will be identified and additional administrative sources may be added to further refine the database, along with possible methodological improvements. An interim report will be available in May 2014 detailing the results of this first phase of investigation.

The second version of the CSDD proposes to incorporate the identified sources and methodological improvements and to reproduce the file as of May 2011. This second phase will gauge the extent to which the database has been improved; the findings will be presented in a second report by March 2015. This report will provide recommendations for future development and suggest possible uses that are commensurate with the quality of the database. Should development and refinements to the database continue, it could ultimately be compared to the 2016 Census of Population to further gauge its potential.

Depending on the degree of success achieved by the CSDD, possible uses might include support for Census Program with regards to planning, non-response follow-up, processing, and quality assessment, as well as support for coverage studies and the population estimates program. Should the quality of the data indicate fitness for use, additional socio-economic variables (e.g. mother tongue, income, education) could be added to make it usable as a sampling frame for traditional household surveys targeting rare populations or as a "stand alone" source of data for analytical purposes. Ultimately, the CSDD could either partially replace the traditional Census in parts of the country where the CSDD coverage is strongest or, if the quality of the database is found to be sufficiently robust on a national level, it might be viewed as a viable alternative to the traditional Census collection.

While it is still far too early in the process to exhaustively list all of the issues and problems that may be encountered in the course of the project, there are a number of known challenges and constraints that could arise.

From a conceptual perspective, the current traditional Census has been a count of the population according to their usual place of residence in Canada at a given moment in time (most recently a date in May or June). Neither of these concepts (place and time) is consistently measured in administrative files. For instance, while not problematic for a large majority (~80%) of Canadian residences in urban areas, many rural residents do not, or cannot, provide a civic address that uniquely describes their physical location on their government forms; instead they supply a mailing address (e.g., a Post Office Box or Rural Route), thus making it difficult to accurately locate the dwelling.

From Statistics Canada's experience, it is clear that emigration will also be problematic, because while entries into the country are registered, exits are not. Similarly, migration within Canada is difficult to capture from administrative sources. Moreover, a long lag often exists between an event (such as a move) and its observation in administrative records.

Administrative sources are not designed to derive the internal relationships within a household. Analytically, the traditional Census allows for the creation of many types of familial clusters by collecting information about the relationships of everyone in the household. While administrative records may be able to reproduce some concepts such as a Census family (parents and children), other family types, such as economic families, are more problematic.

In order for the CSDD to be viable in the long term, it is important that the sources for the database remain stable. Statistics Canada would have to continue receiving the data in a timely fashion and the data must remain conceptually consistent.

Finally, should the viability of the CSDD be demonstrated, work on other legacy processes will be required. For instance, the current coverage studies use administrative data to validate the Census of Population. Similarly, the national, provincial and territorial population estimates rely on administrative records for updates. Since administrative records would be used in the creation of the CSDD, such programs would have to rethink their methodology (e.g. survey-based coverage studies). Likewise, a variety of corrective processes have incrementally developed around the traditional Census. Such adjustments would need to be modified to be of use to the CSDD as a partial or total replacement for the Census.

Conclusion

In the short time that the ADS has been in place, much work has been started related to the use of administrative data at Statistics Canada. In addition to the initiatives reported above, the ADS, in collaboration with the Methodology Branch of Statistics Canada, has started the development of an evaluation framework to assess the quality of potential administrative data sources and their statistical usability. To aid in the development of this framework, the ADS is reviewing what exists internationally, including the Data Quality Assessment Tool for Administrative Data developed in the United States. This framework should help meet three objectives:

1. To determine the statistical usability of the administrative datasets
2. To help make good corporate decisions about which administrative datasets to acquire
3. To base the acquisition on a formal process that demonstrates that administrative datasets are acquired by Statistics Canada to fulfill its mandate

Finally, given the increasing interest in Big Data, the Administrative Data Secretariat has established a Big Data community of practices with the following objectives: to gain knowledge and share experiences; to engage with colleagues internally or externally when needed; and to report findings to senior managers when appropriate. The community meets once a month to share experiences. During these meetings the community identifies issues or challenges of common interest and may mandate one or more members of the community to examine it more closely and report to the community. To date, one of the topics discussed that led to the creation of a working group was the procurement process to acquire administrative data from the private sector.

References

Royce, D. 2013. *A Survey of International Frameworks for the Statistical Use of Administrative Data*. Administrative Data Secretariat. Statistics Canada.